TECHNISCHE
UNIVERSITÄT
DARMSTADT

Monocular Depth Estimation Using Atrous Convolutions

Supervisor: Faraz Saeedan

Group 5

Fabian Kessler, Dominik Straub, Steven Lang

February 18, 2019

Introduction

Experiments & Results

Discussion

Conclusion

Monocular Depth Estimation

- Monodepth: Estimate depth from a single image at test-time (instead of stereo pair)
- Existing approaches treat this as a supervised regression problem
- Godard et al., CVPR 2017: Depth estimation as a stereo reconstruction problem

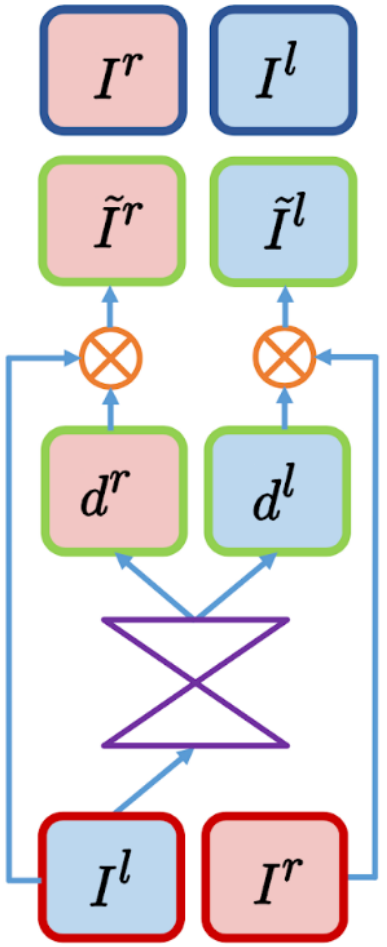


Figure: Godard et al. 2017

Atrous Convolutions

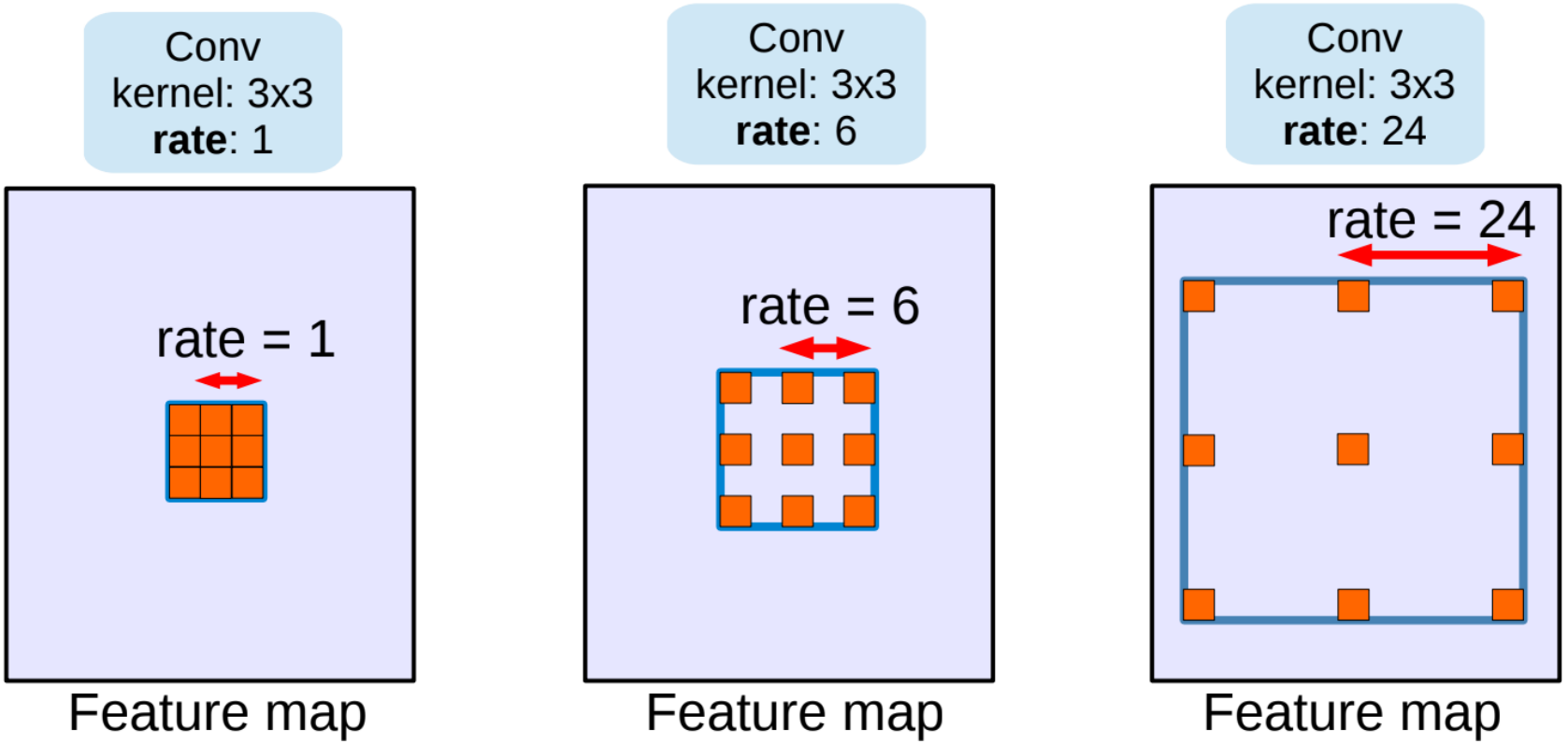
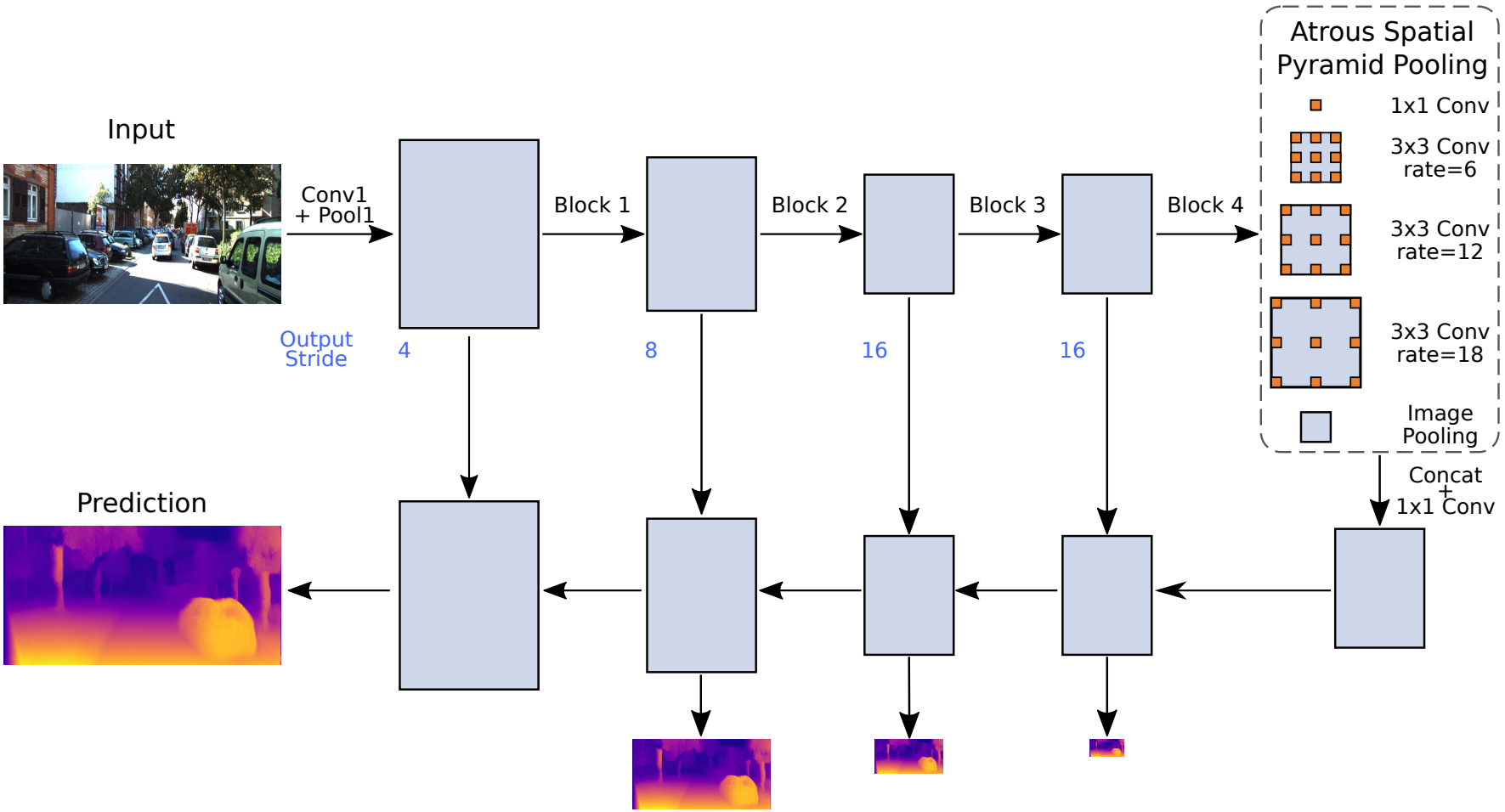


Figure: Chen et al., 2017

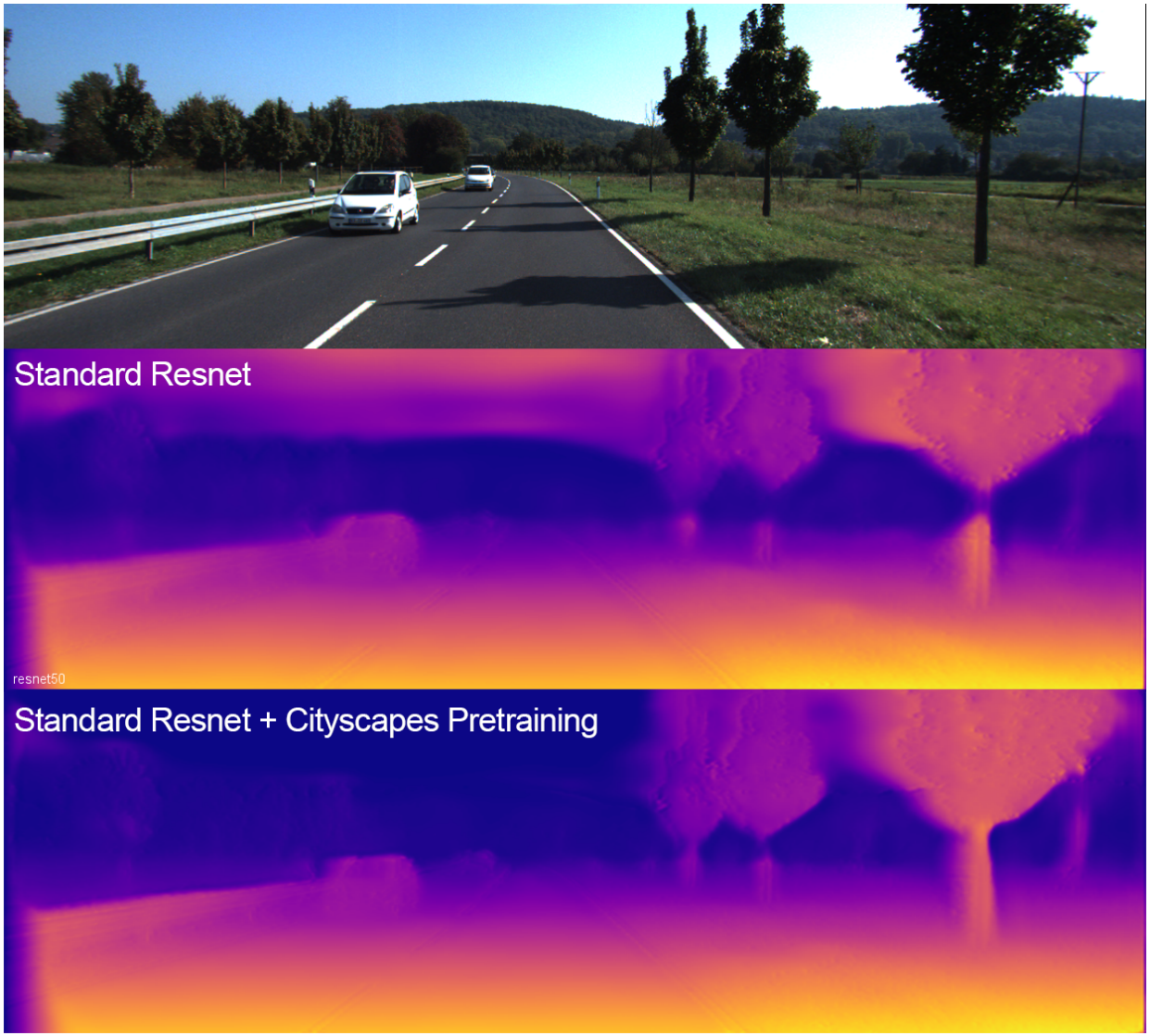
Project Objective

- Get Monodepth baseline to run
- Design and implement encoder-decoder network using atrous convolutions
- Consider memory and runtime in architectural decisions
- Quantitative comparison of proposed architectures to baseline

Our Architecture with ASPP



Remember last time? Sky Artifacts



How we solved it: Reimplementation

mrharicot commented on Dec 8, 2017 Owner

Hi **@7LFB**,
I think you are correct.
My implementation of Resnet50 was quite last minute and has actually another bug.
I will leave it as is for now to make sure people can reproduce the results from the publication.
I however plan on having either a branch or a fork to fix all the bugs at some point in the near future.

How we solved it: Reimplementation

mrharicot commented on Dec 8, 2017 Owner

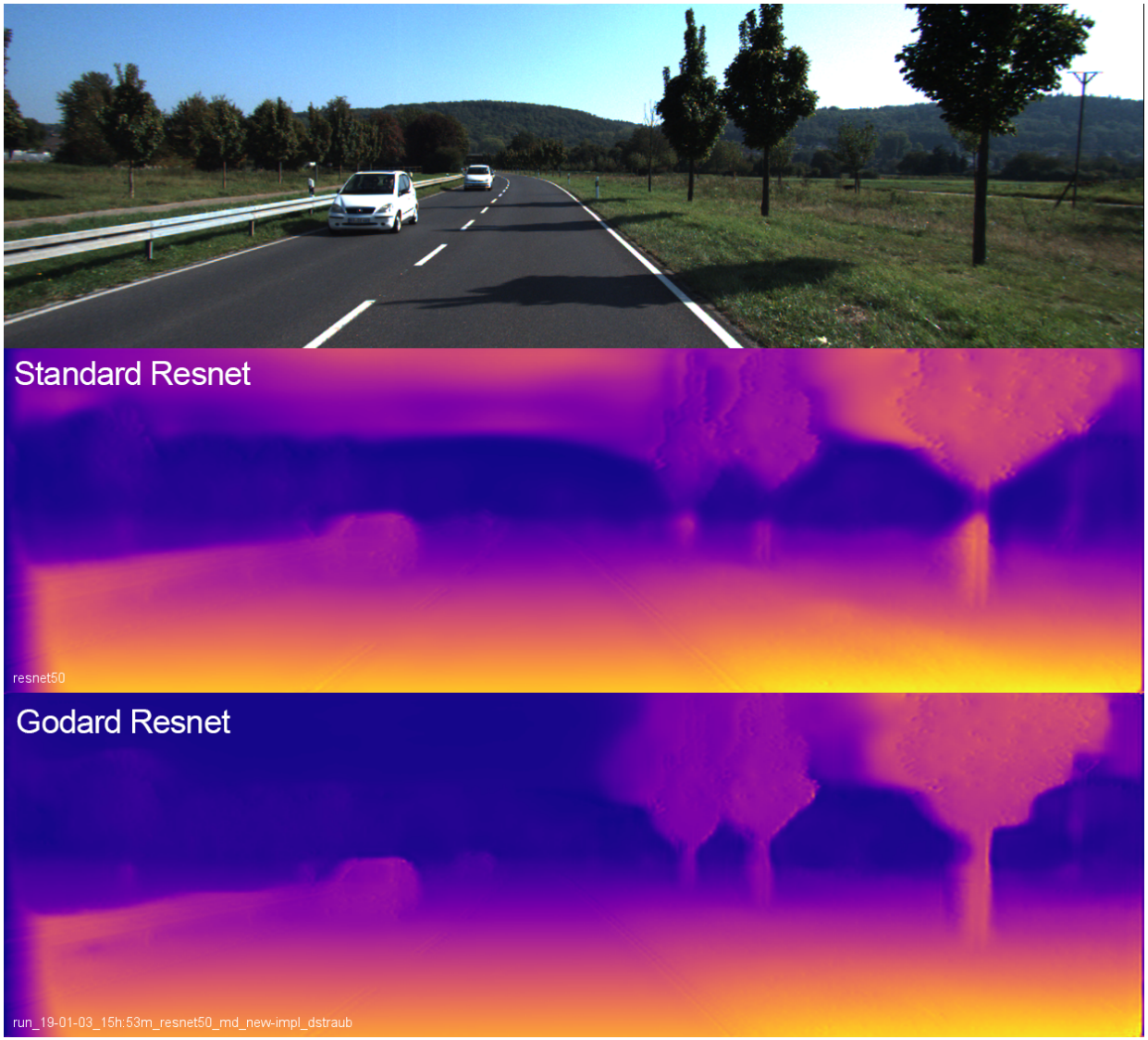
Hi **@7LFB**,
I think you are correct.
My implementation of Resnet50 was quite last minute and has actually another bug.
I will leave it as is for now to make sure people can reproduce the results from the publication.
I however plan on having either a branch or a fork to fix all the bugs at some point in the near future.

Differences of their implementation compared to standard ResNet:

- No batch-normalization
- Nearest-neighbor instead of bilinear interpolation
- Order of convolution strides in ResNet switched

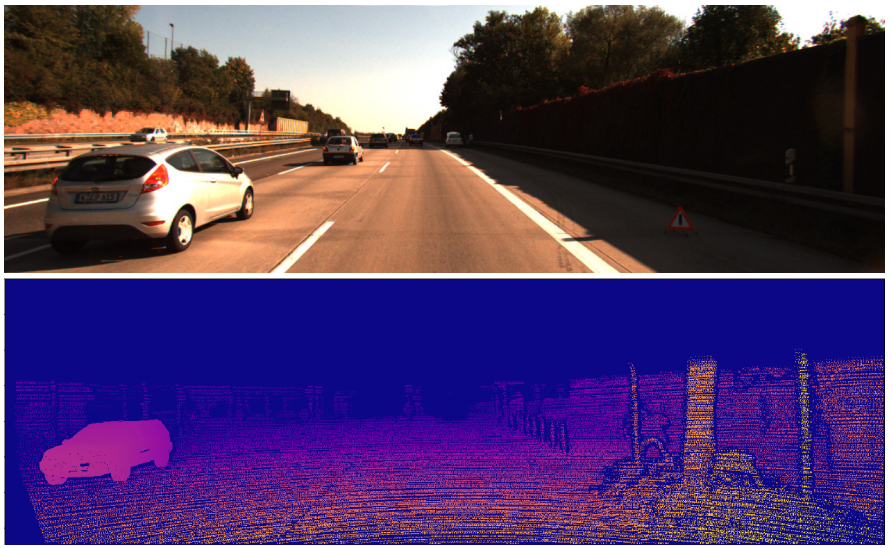
\+ Fixed bug in author's implementation of loss function

New Disparity Maps



General Experimental Setup

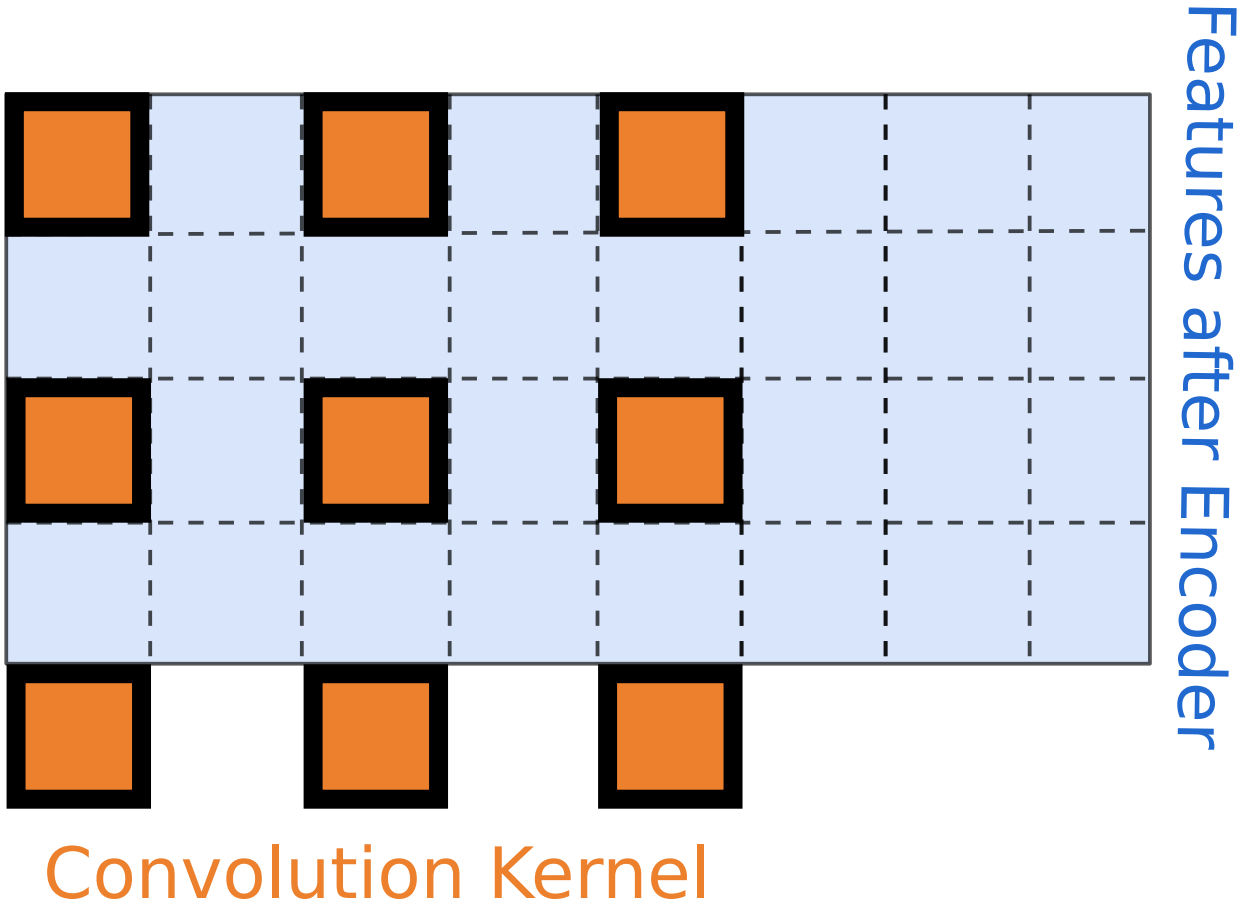
- General setup
  - Batch Size 16
  - Learning Rate 2e-4
  - 50 epochs
- KITTI for training
  - Rescaled to 256 x 512 px
- KITTI Stereo 2015 for testing
  - Sparse ground truth



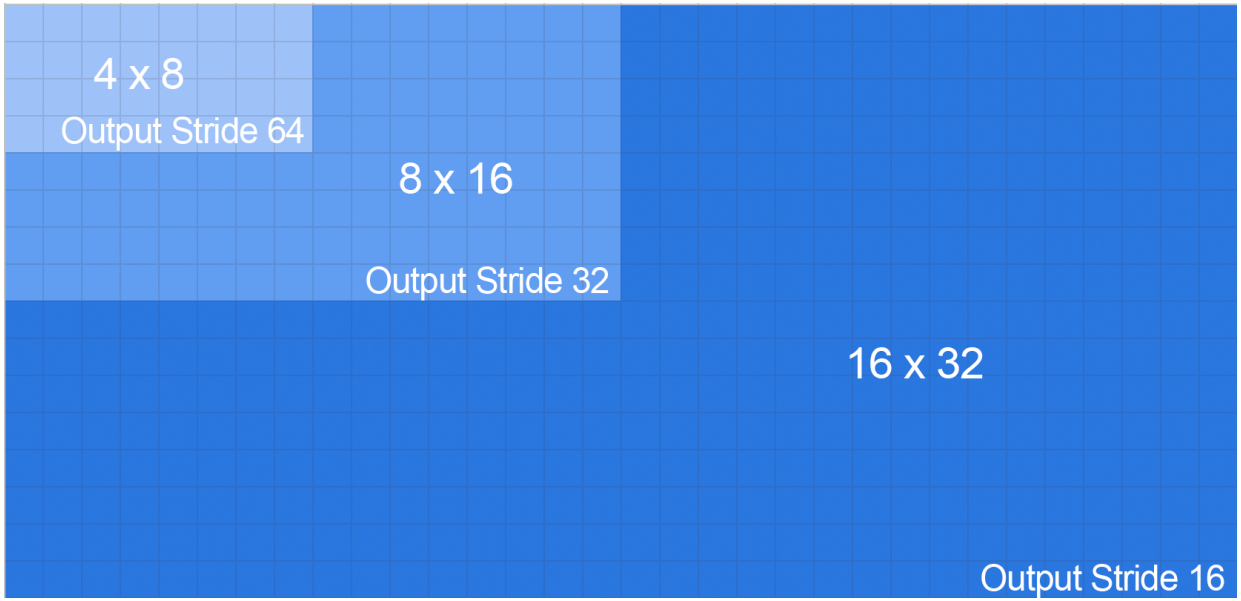
Experiments: Output Strides

Issues

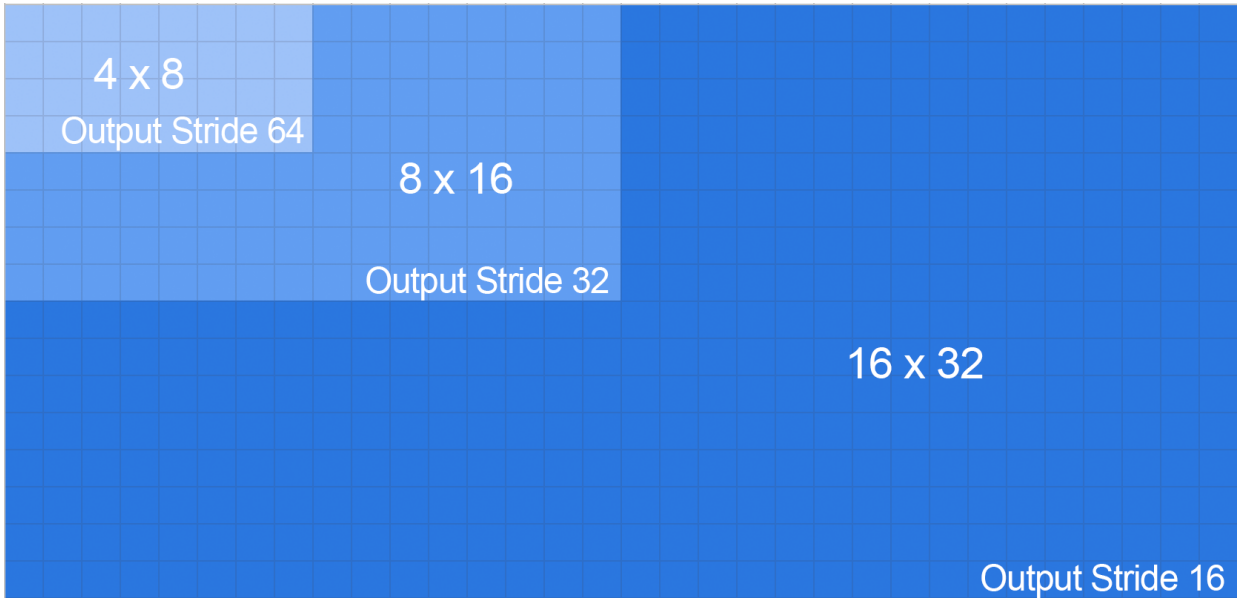
Output stride 64 with Atrous rate > 1



Experiments: Output Strides

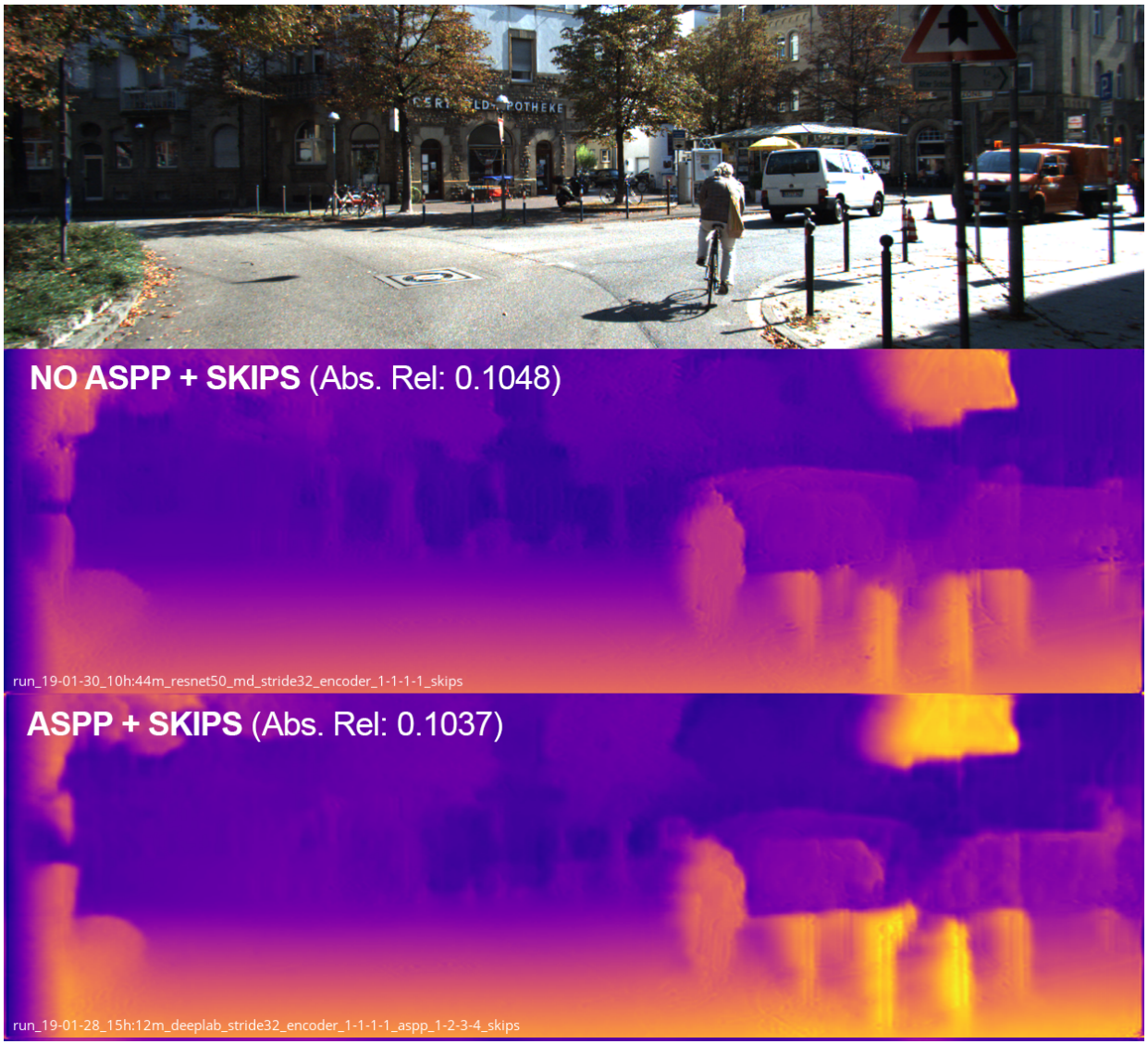


Experiments: Output Strides

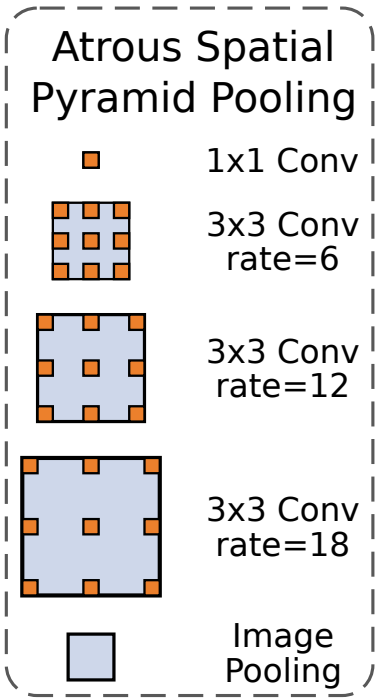


Stride	Abs. Rel.	Runtime $\left(\frac{Min.}{Epoch}\right)$	Memory (GB)	Param. (M)
64	0.1120	**15**	**6.12**	58.4
32	0.1048	29	8.87	58.4
16	**0.1041**	29	8.88	58.4
8	0.1068	43	10.61	58.4

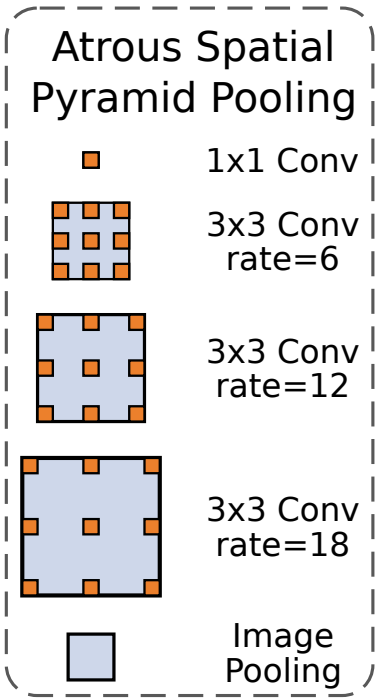
Experiments: ASPP vs. no ASPP

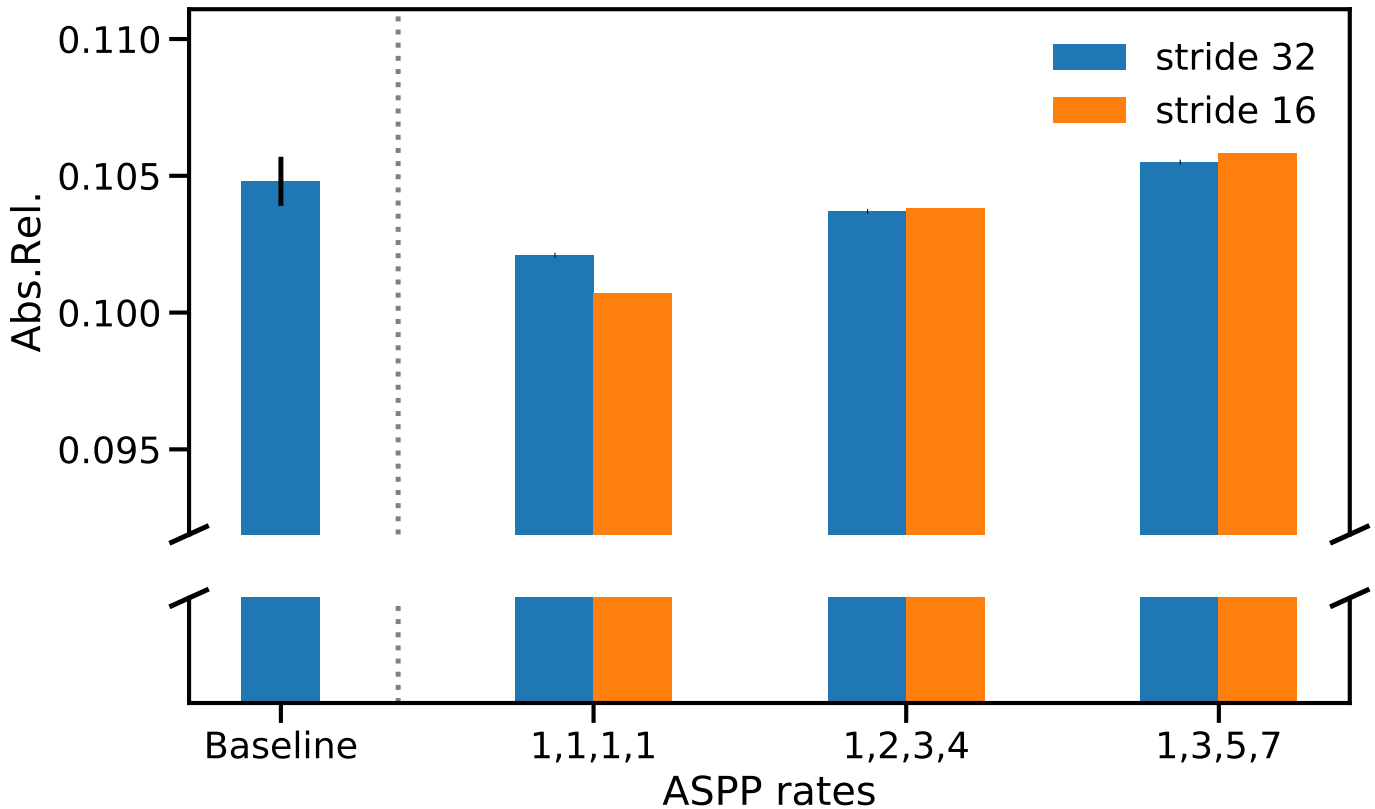


Experiments: Atrous Rates

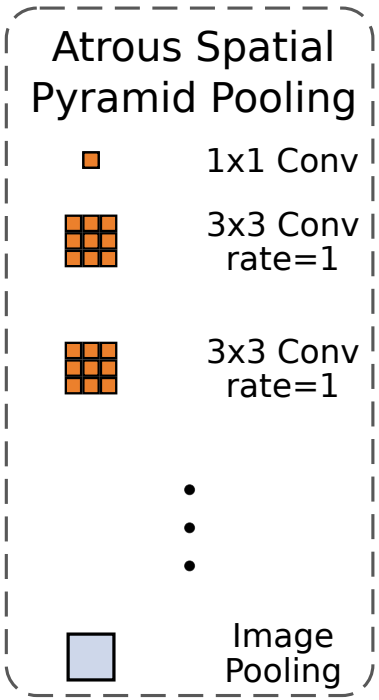


Experiments: Atrous Rates

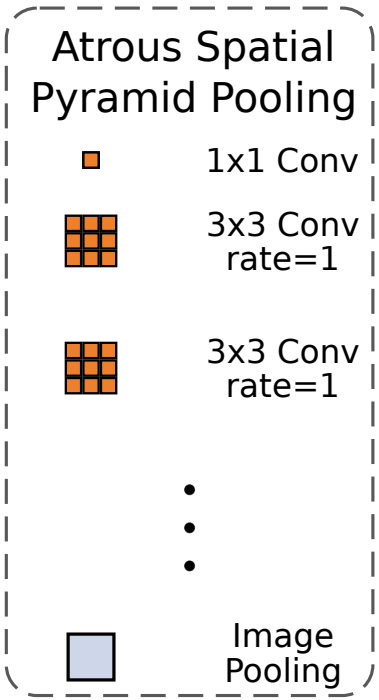


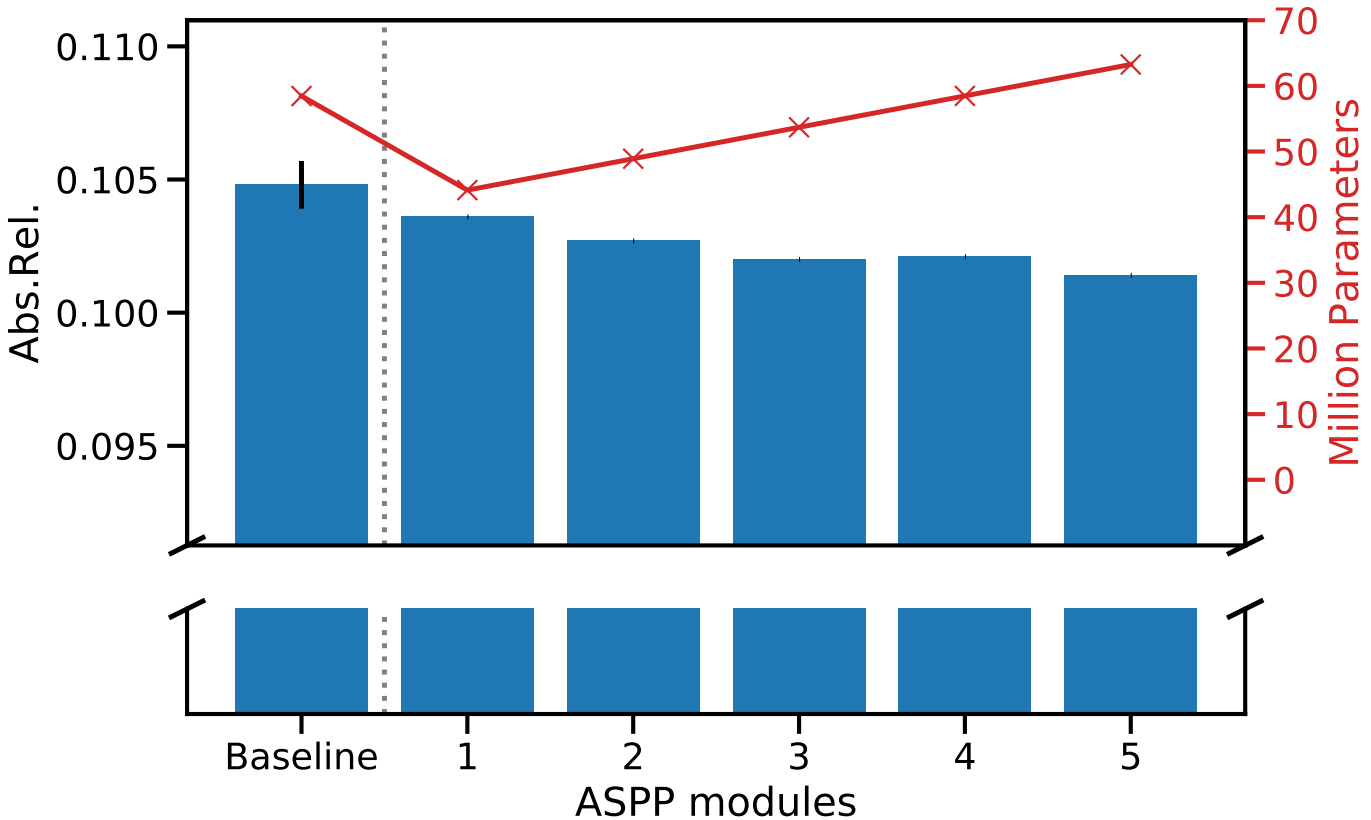


Experiments: Different Number of Modules

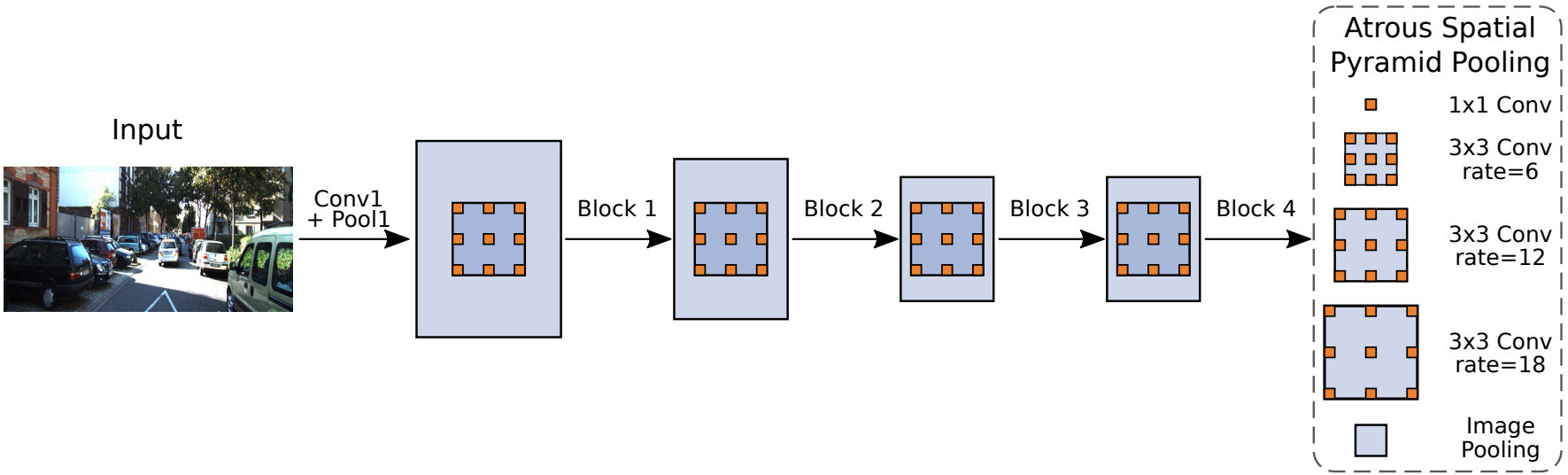


Experiments: Different Number of Modules



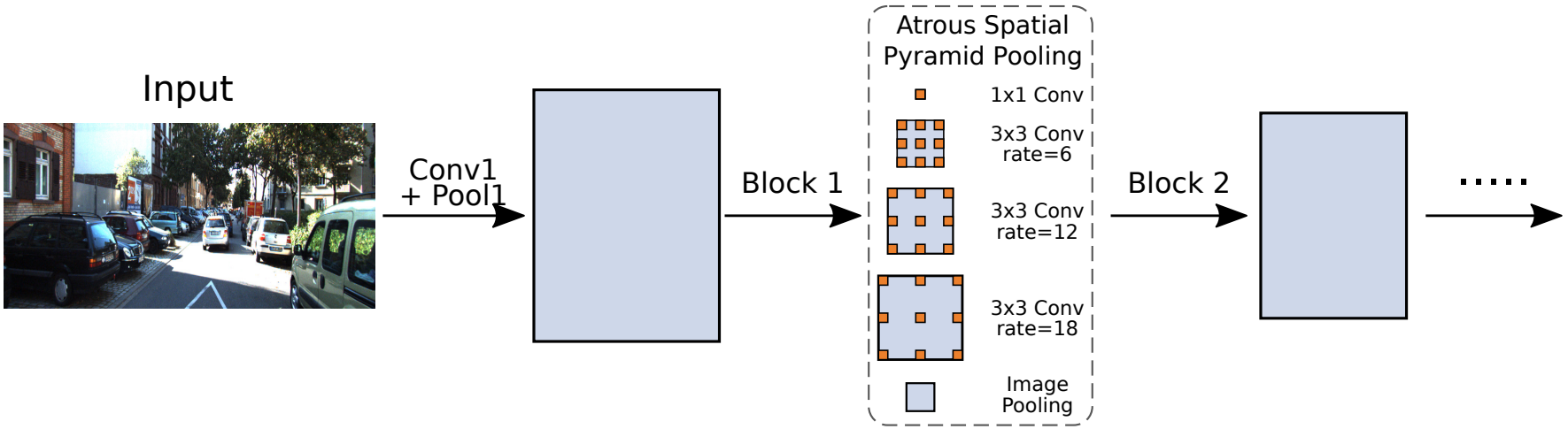


Experiments: Atrous Convolutions in Encoder



No improvement with atrous convolutions in ResNet blocks

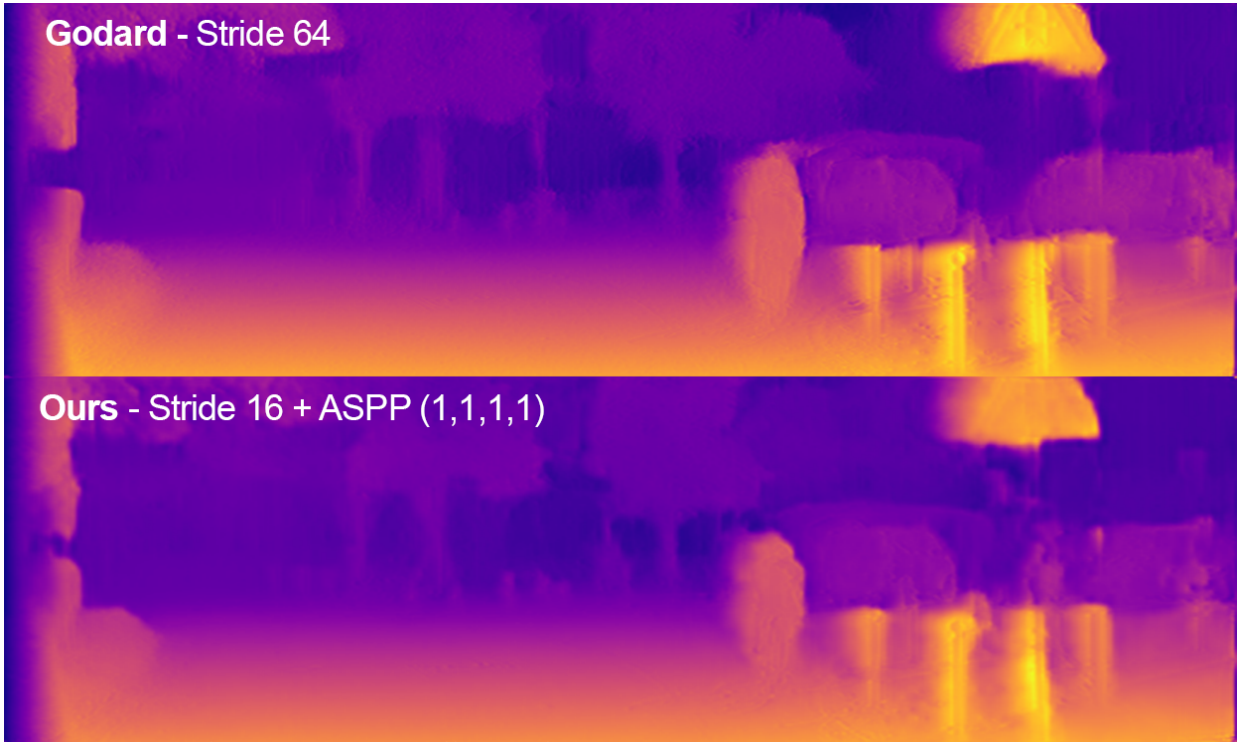
Experiments: ASPP in ResNet



No improvement with ASPP module between ResNet blocks

Final Results

Pretrained on CityScapes & finetuned on KITTI



Output-stride	ASPP	Abs.Rel.	Sq.Rel.	RMSE	Log RMSE	a1	a2	a3	Params (M)	Δ Abs. Rel. (%)
64 - Godard	-	0.0970	0.8960	5.093	0.176	0.962	0.962	0.986	58.4	–
16 - Godard	-	0.0955	0.8752	4.937	0.174	0.881	0.961	0.984	58.4	2.57
16 - Ours	1-1-1-1	**0.0927**	**0.8132**	**4.865**	**0.168**	**0.888**	**0.967**	**0.987**	58.4	**4.43**
32 - Ours	1	0.0941	0.8196	4.910	0.173	0.882	0.963	0.986	**44.1**	2.99
32 - Ours	1-2-3-4	0.0936	0.8281	4.941	0.172	0.884	0.963	**0.987**	58.4	3.50

Discussion

- Hyperparameter tuning
  - Learning rate
  - Loss hyperparameters

$$C_s = \boldsymbol{\alpha_{ap}} \left(C_{ap}^l + C_{ap}^r\right) + \boldsymbol{\alpha_{ds}} \left(C_{ds}^l + C_{ds}^r\right) + \boldsymbol{\alpha_{lr}} \left(C_{lr}^l + C_{lr}^r\right)$$

  - Focus on architectural design instead

Discussion

- Output stride 8
- Decoder architecture
- Robustness of training & predictions

Recap

- Reproduced baseline
- Experiments
  - Output stride
  - ASPP rates
  - Number of ASPP modules
  - Atrous convolutions in encoder

Conclusion

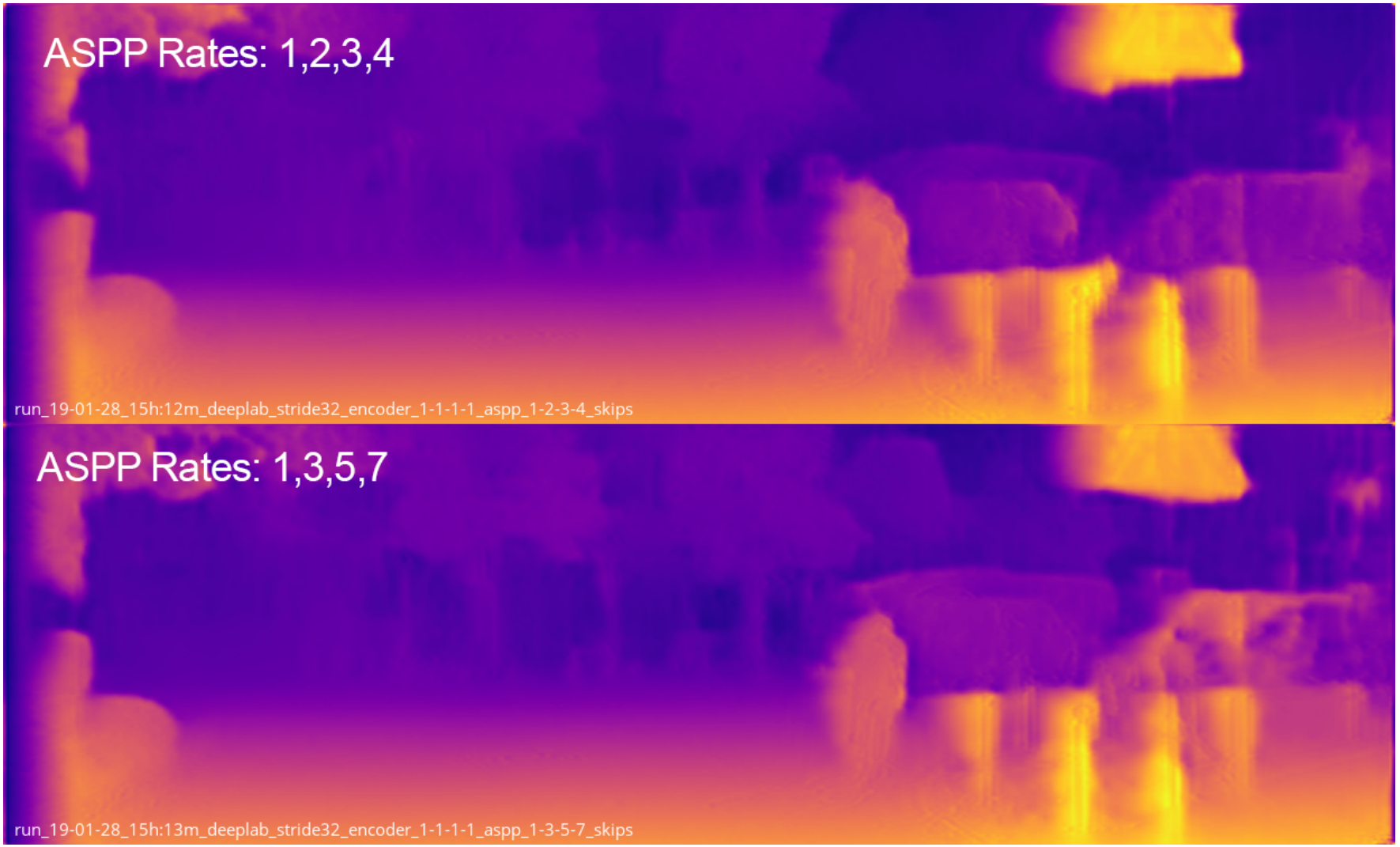
- Atrous convolutions . . .
  - do not improve monocular depth estimation
  - need a decreased output stride, which harms runtime and memory
- Channel reduction after encoder . . .
  - decreases parameter count and improves runtime
  - without losing predictive power

References

- C. Godard, O. Aodha and G. Brostow: Unsupervised Monocular Depth Estimation with Left- Right Consistency, CVPR 2017.
- L. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv 2018.
- M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele: The cityscapes dataset for semantic urban scene understanding. CVPR. 2016.
- M. Menze and A. Geiger: Object Scene Flow for Autonomous Vehicles. CVPR. 2015.

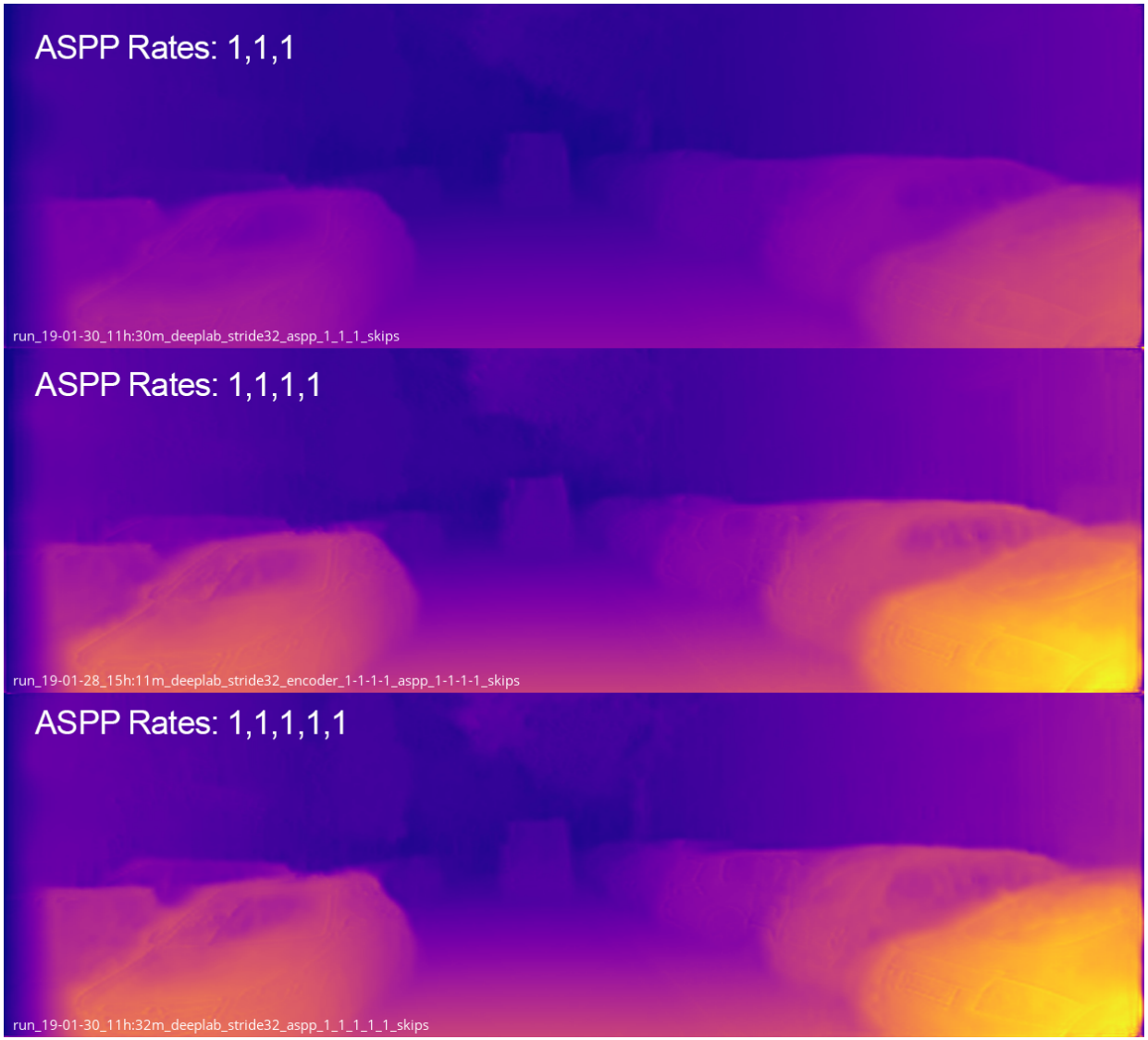
Experiments: Atrous Rates (Visual Comparison)

Experiments: Atrous Rates (Visual Comparison)

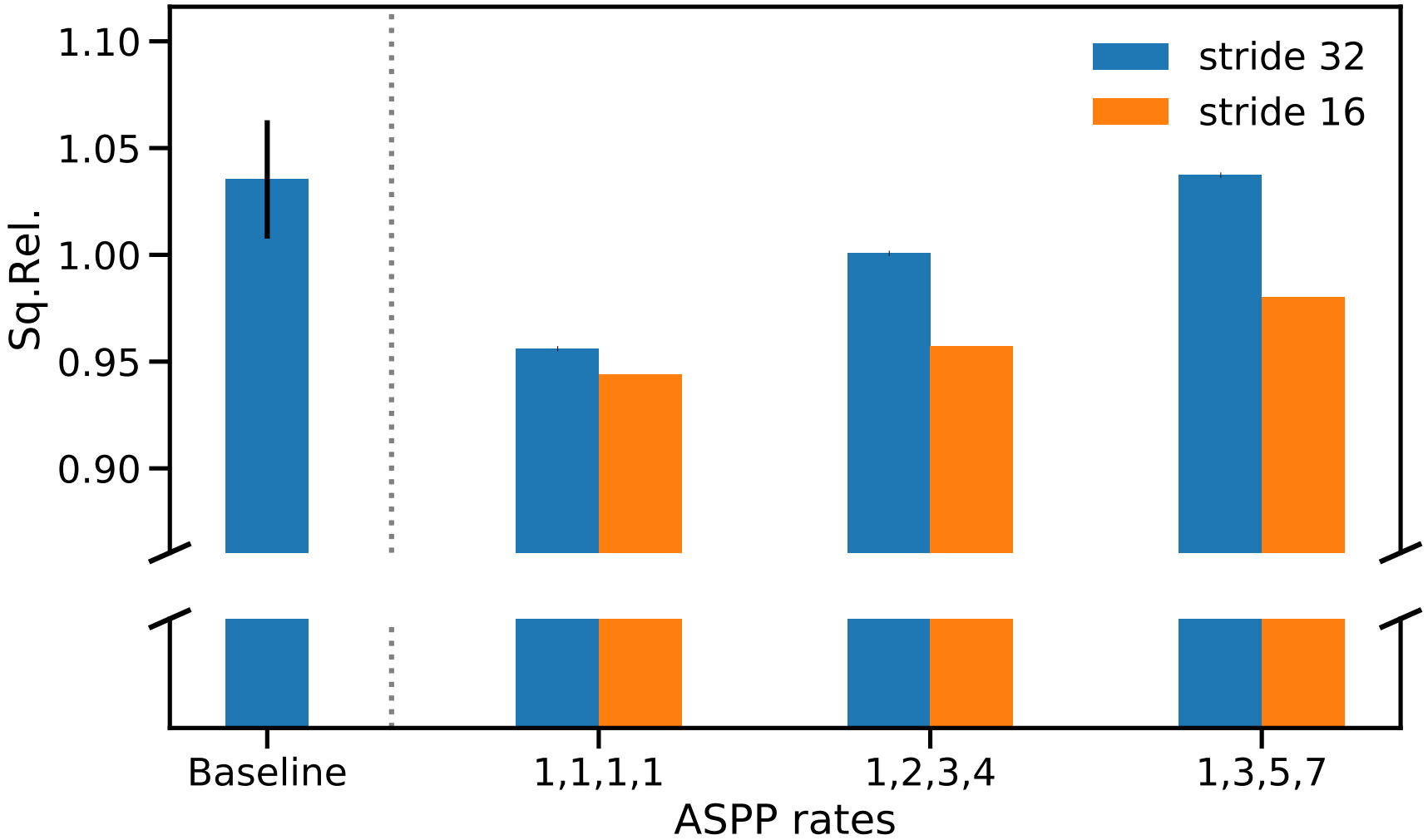


Experiments: Atrous Rates (Visual Comparison)

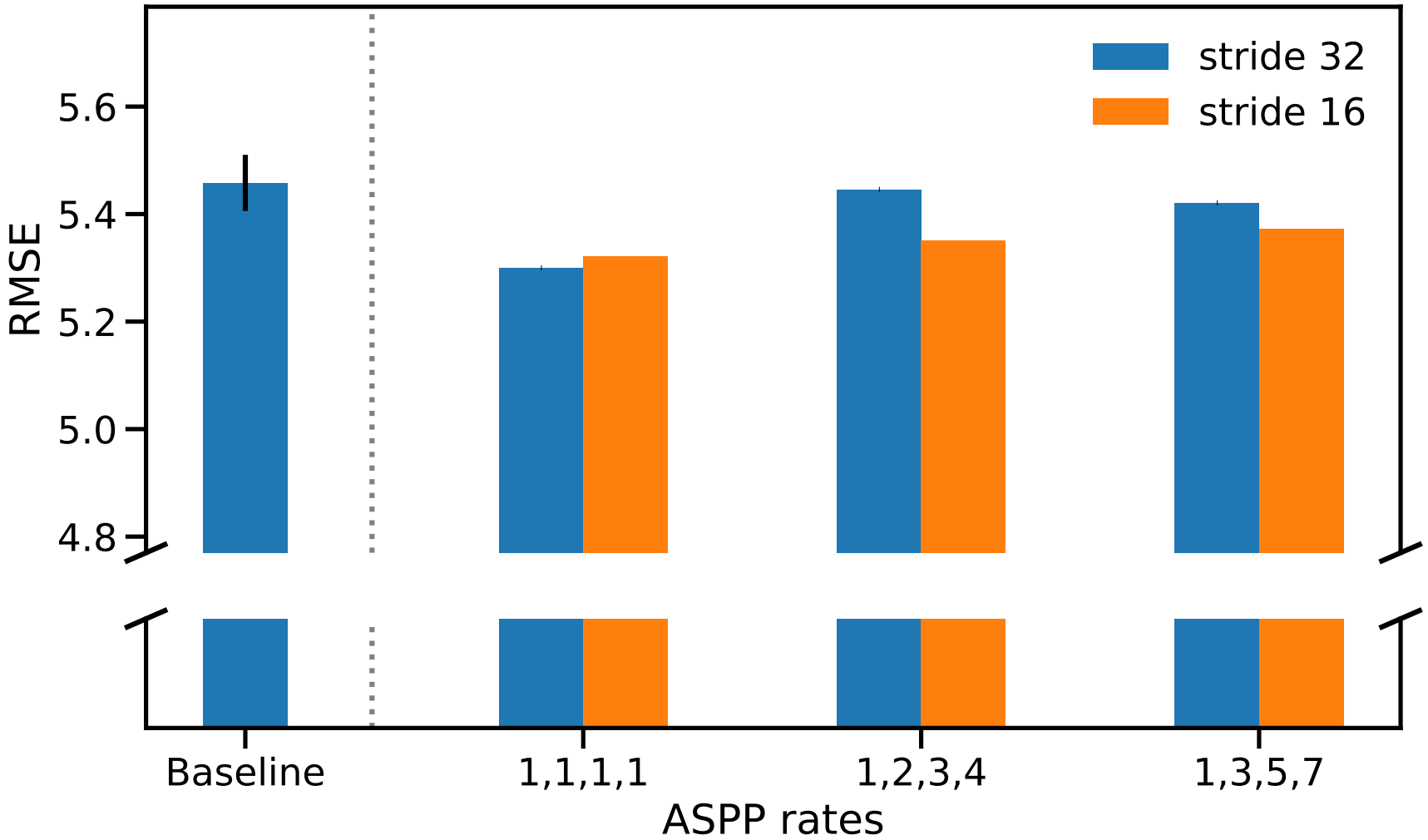
Experiments: Atrous Rates (Visual Comparison)



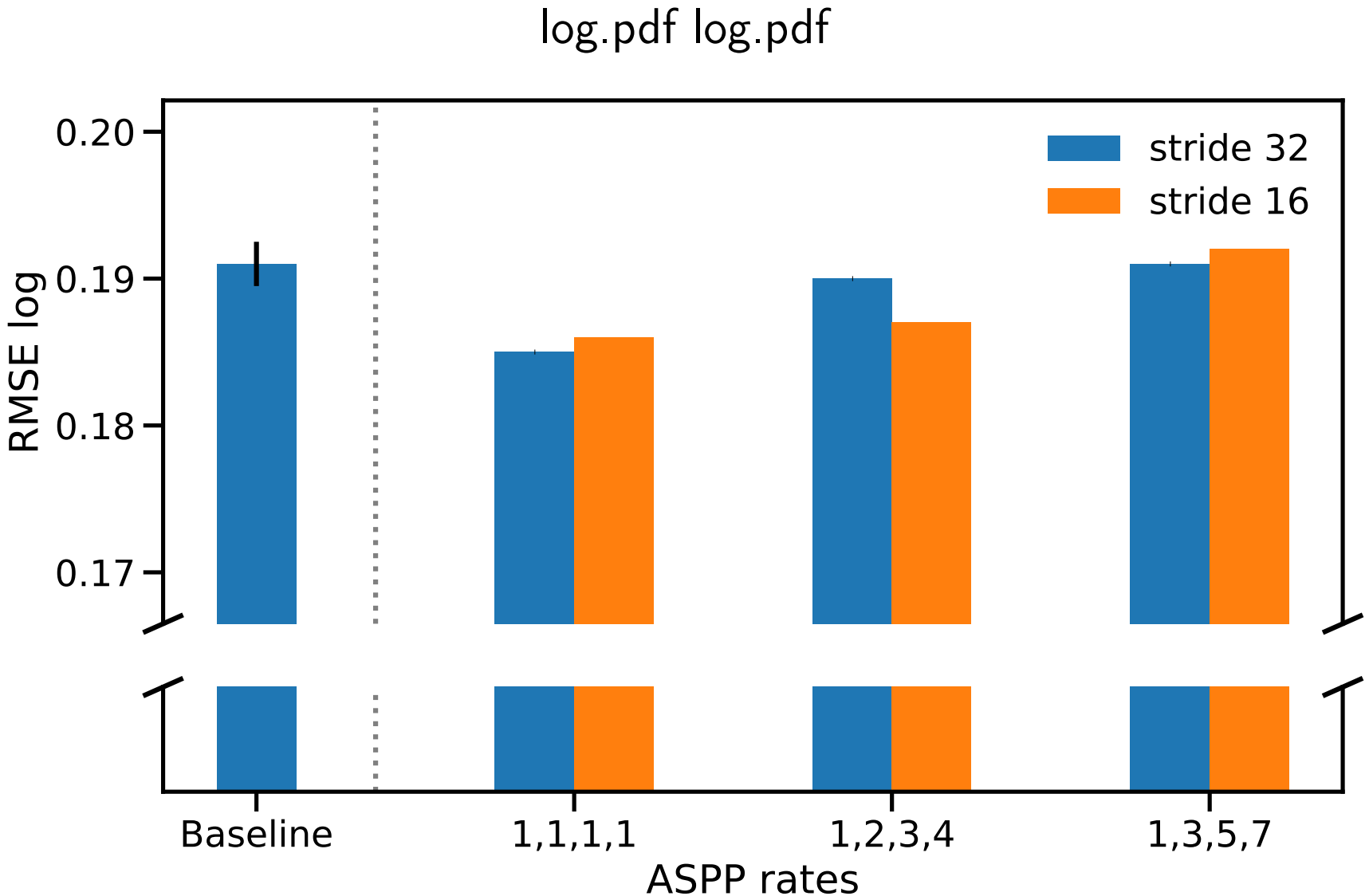
Experiments: Atrous Rates (Squared relative error)



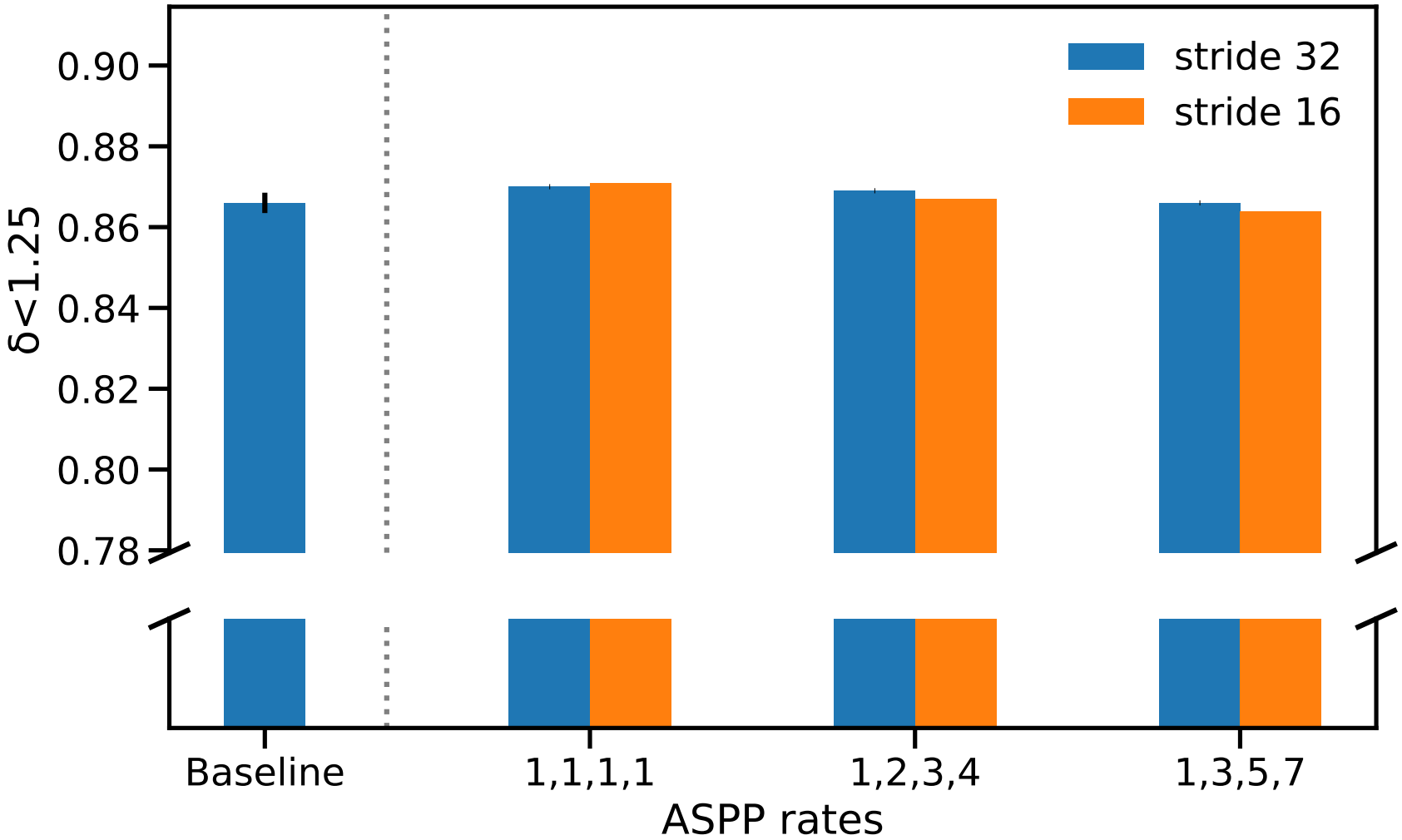
Experiments: Atrous Rates (RMSE)



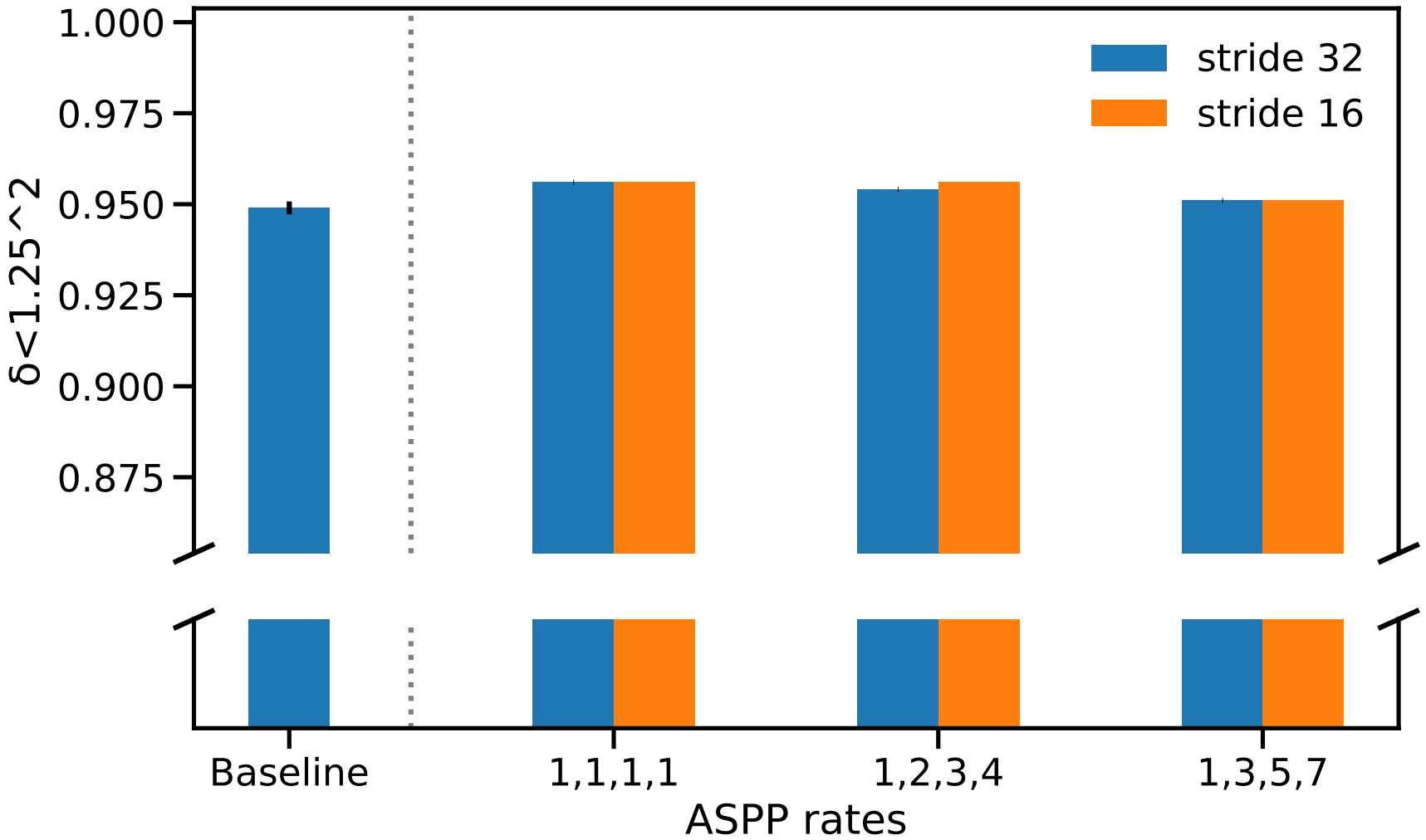
Experiments: Atrous Rates (RMSE of log)



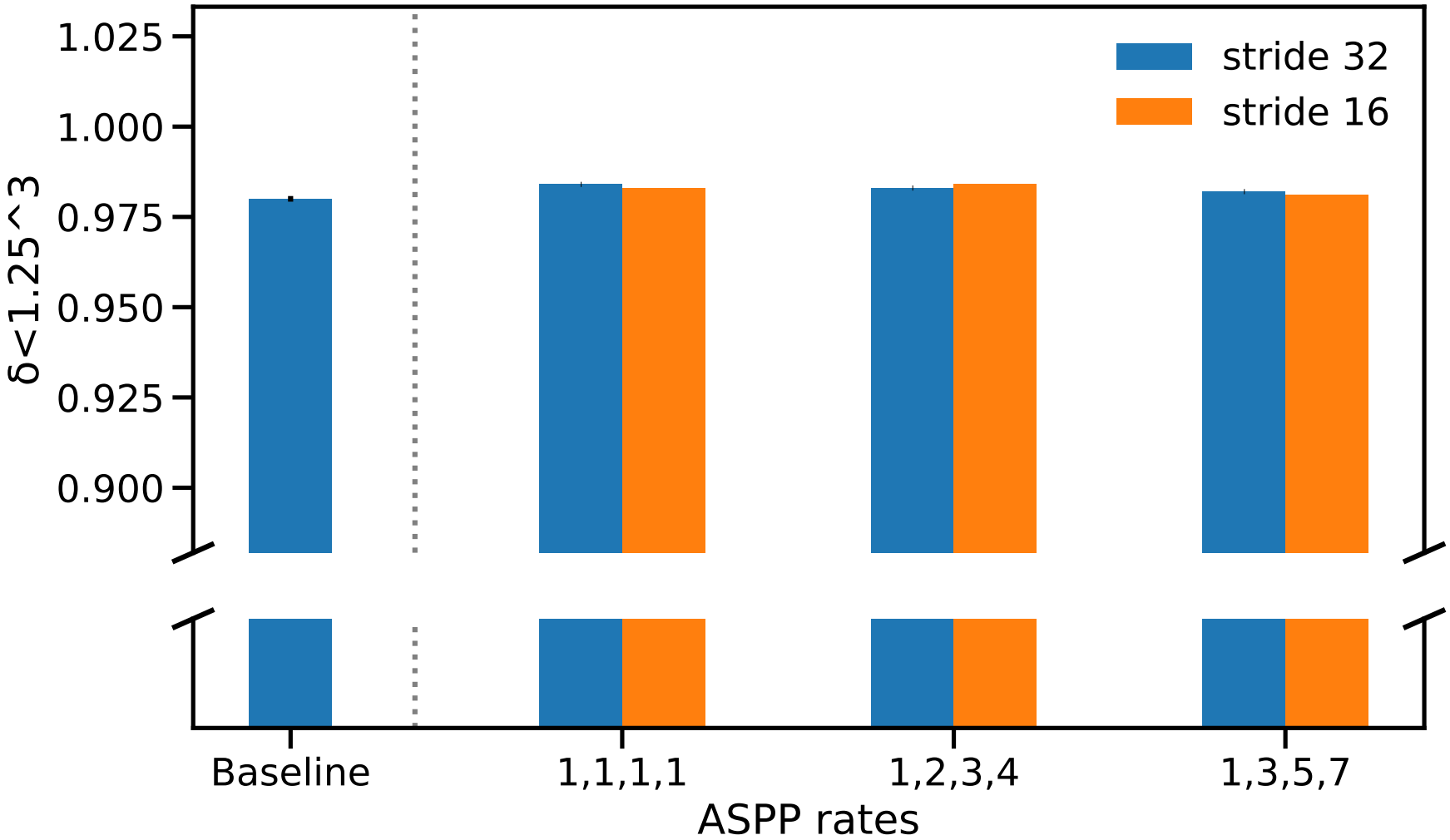
Experiments: Atrous Rates (% inliers 1)



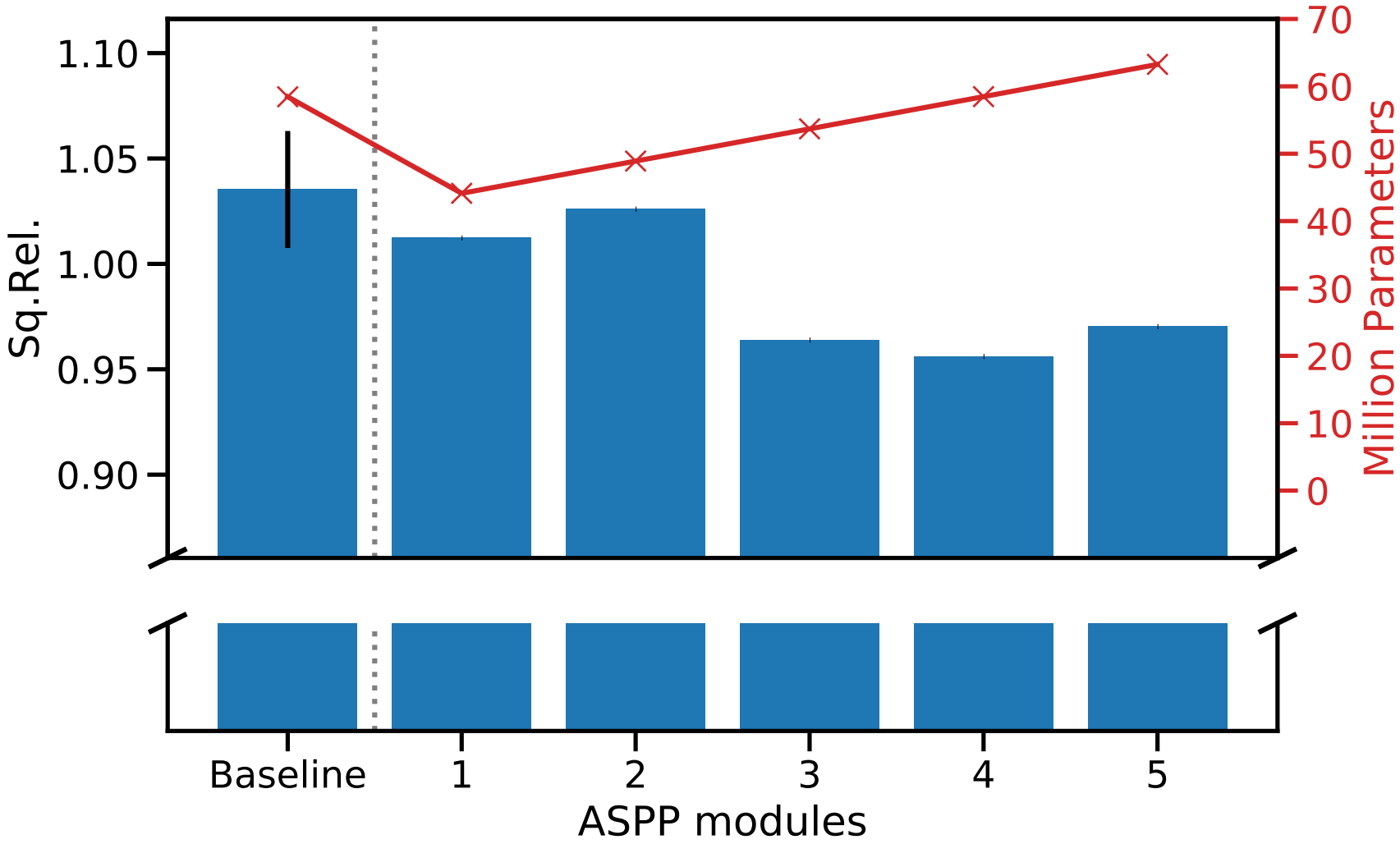
Experiments: Atrous Rates (% inliers 2)



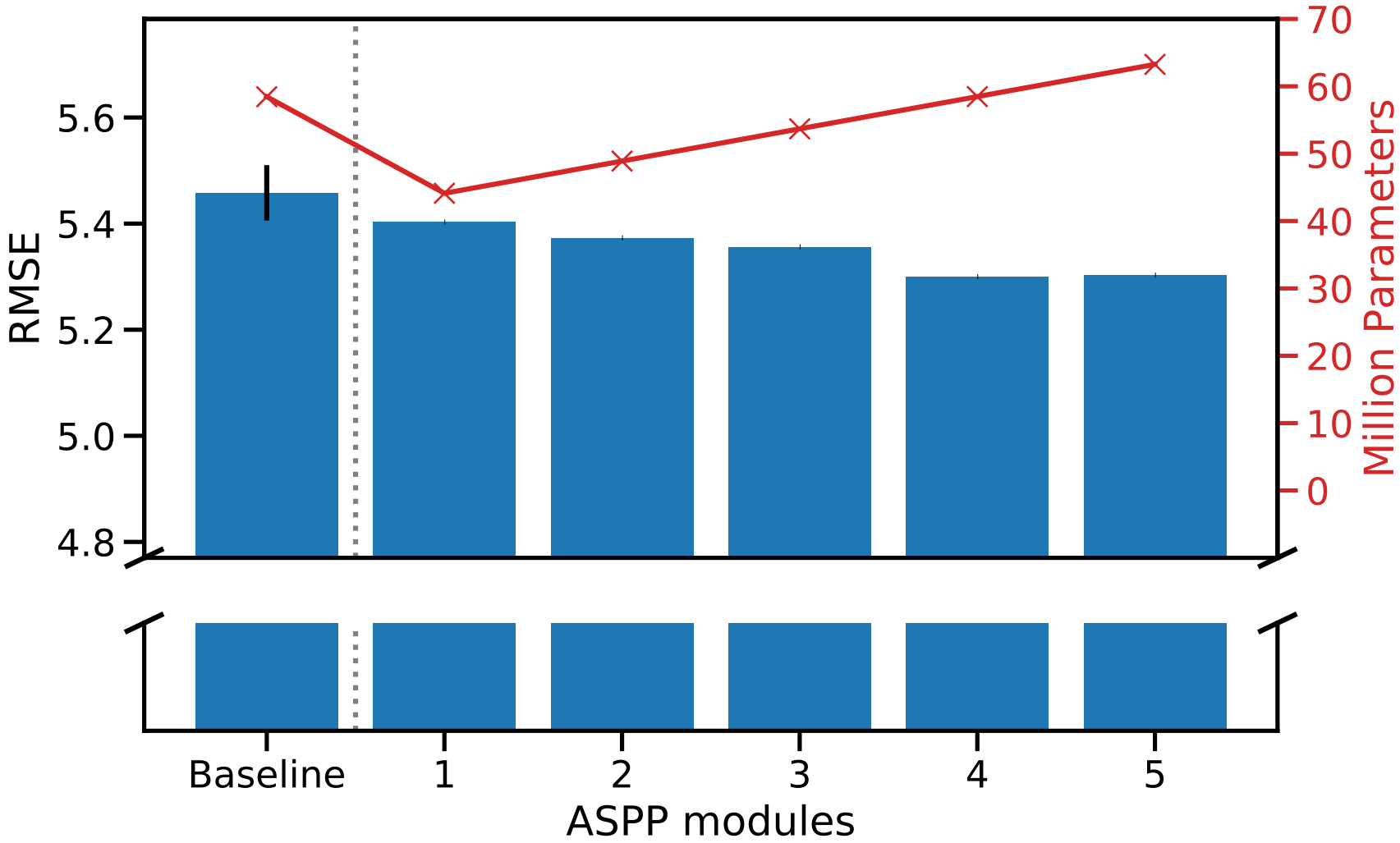
Experiments: Atrous Rates (% inliers 3)



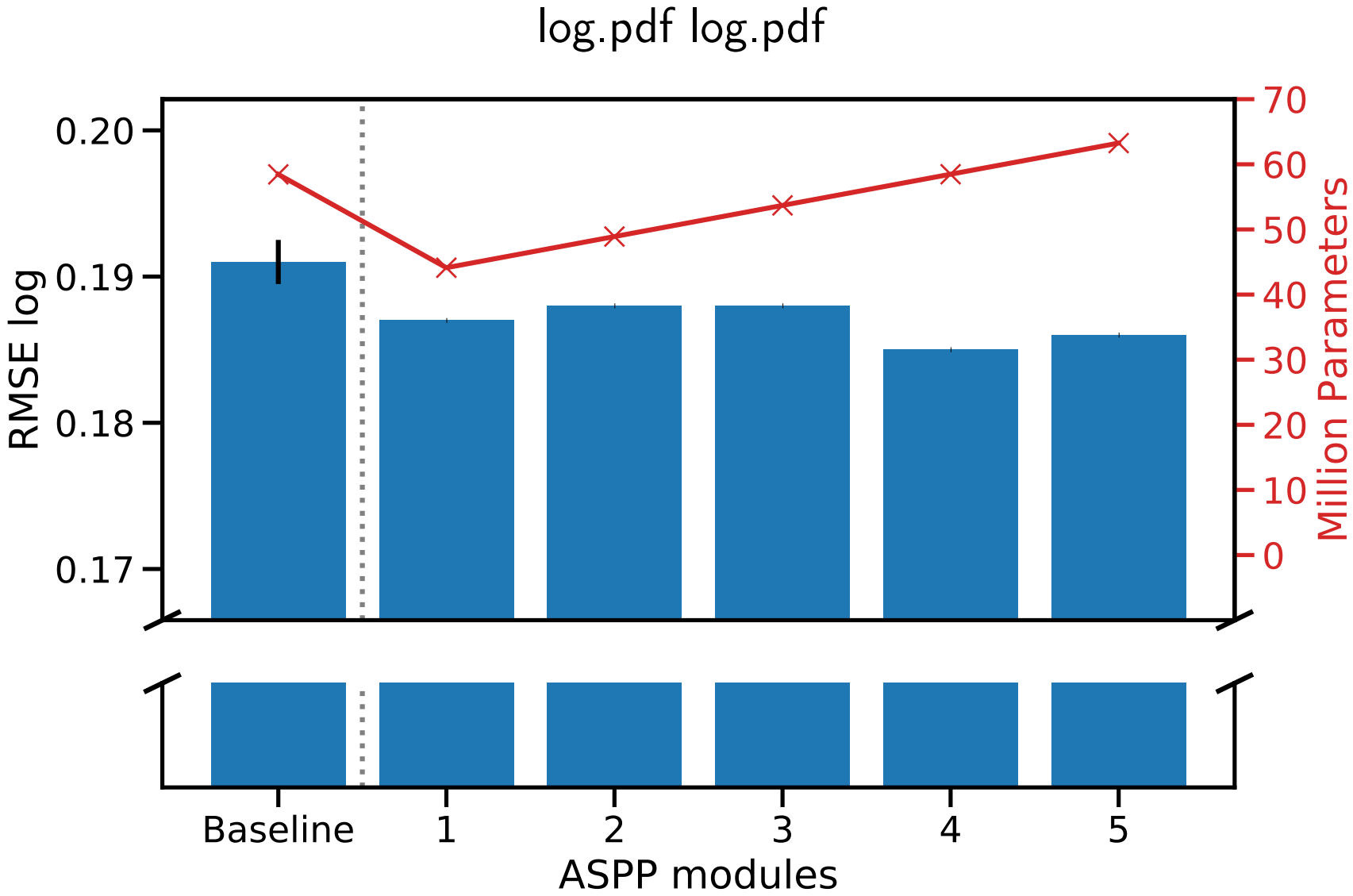
Experiments: Atrous Rates (Squared relative error)



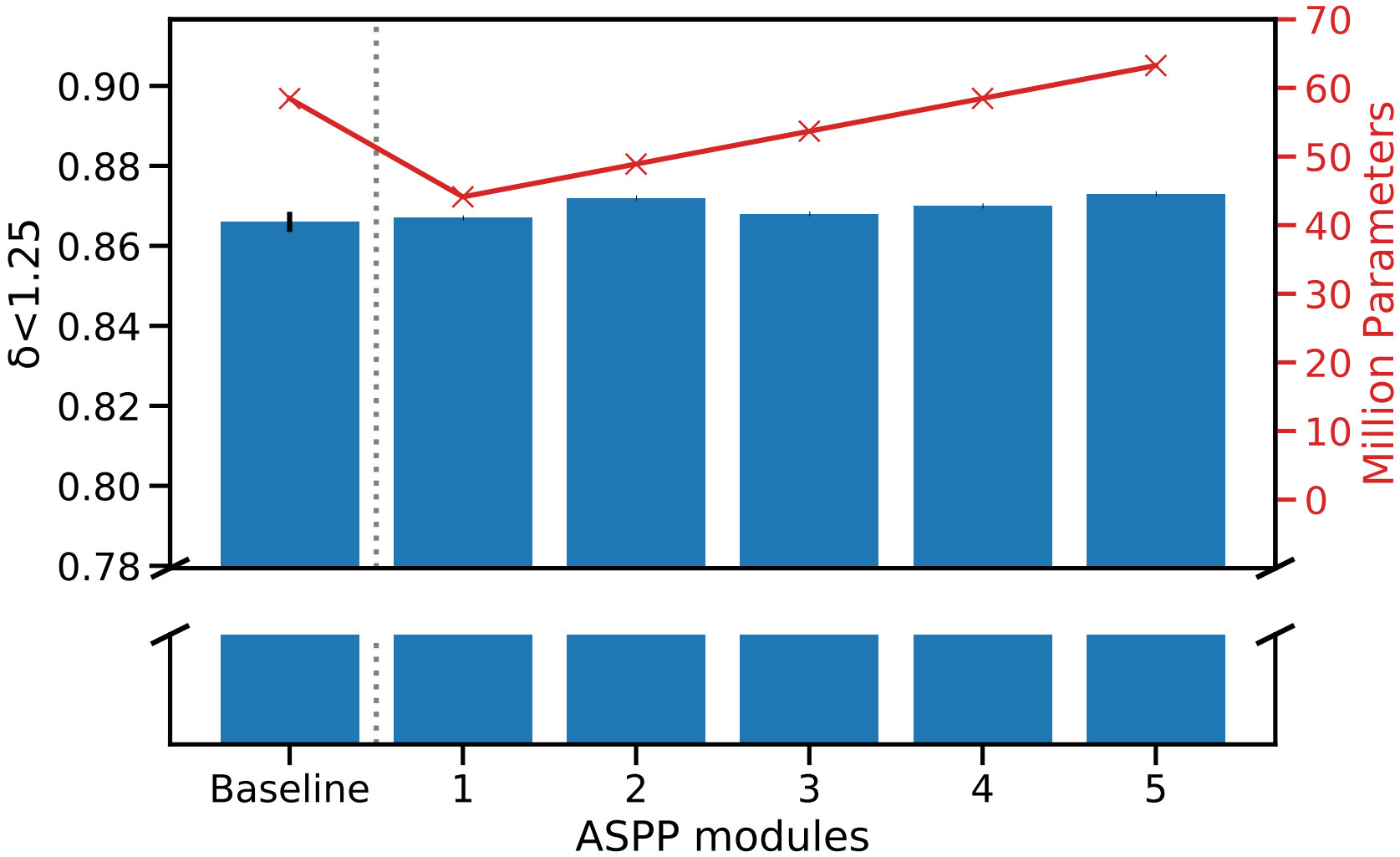
Experiments: Atrous Rates (RMSE)



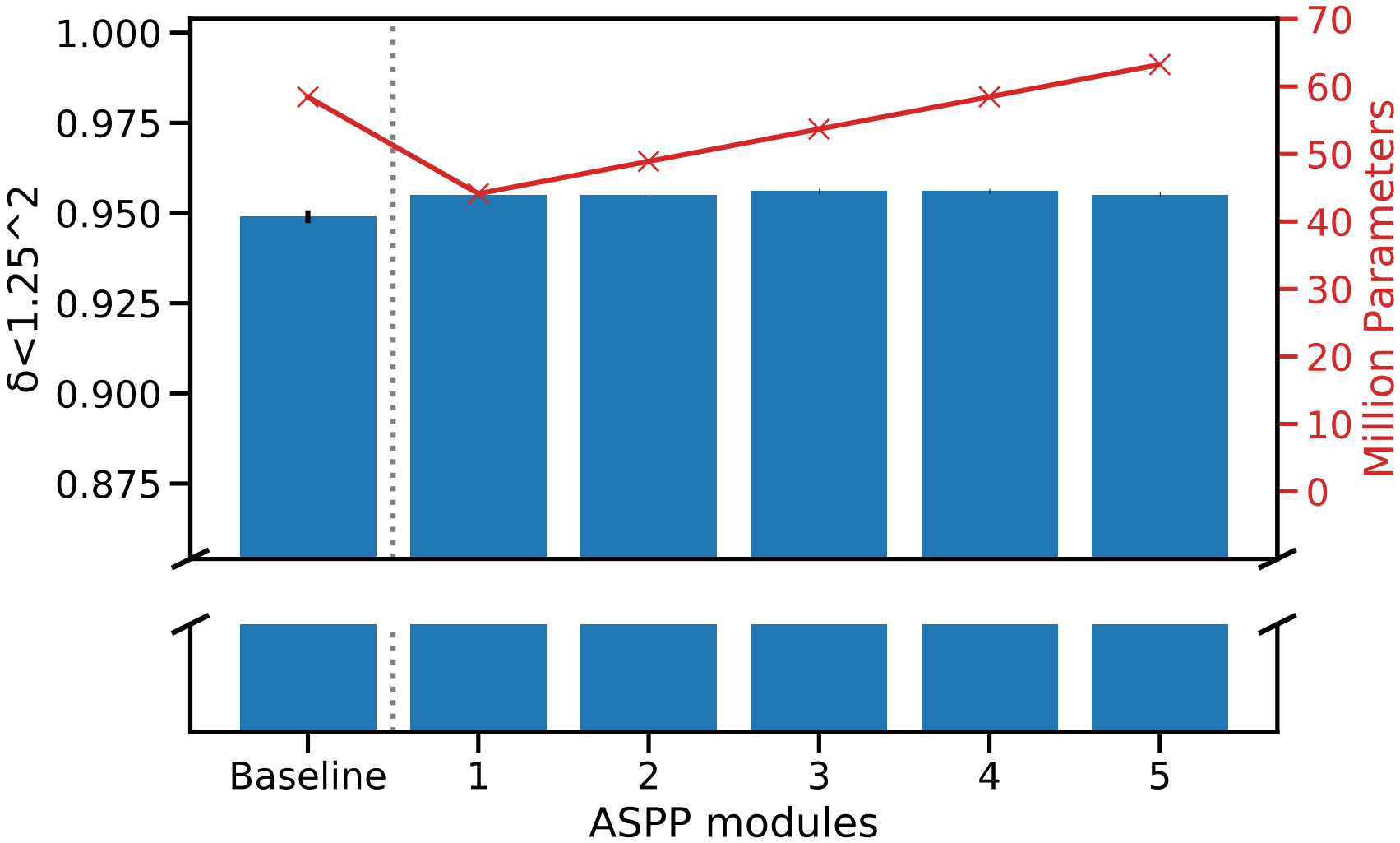
Experiments: Atrous Rates (RMSE of log)



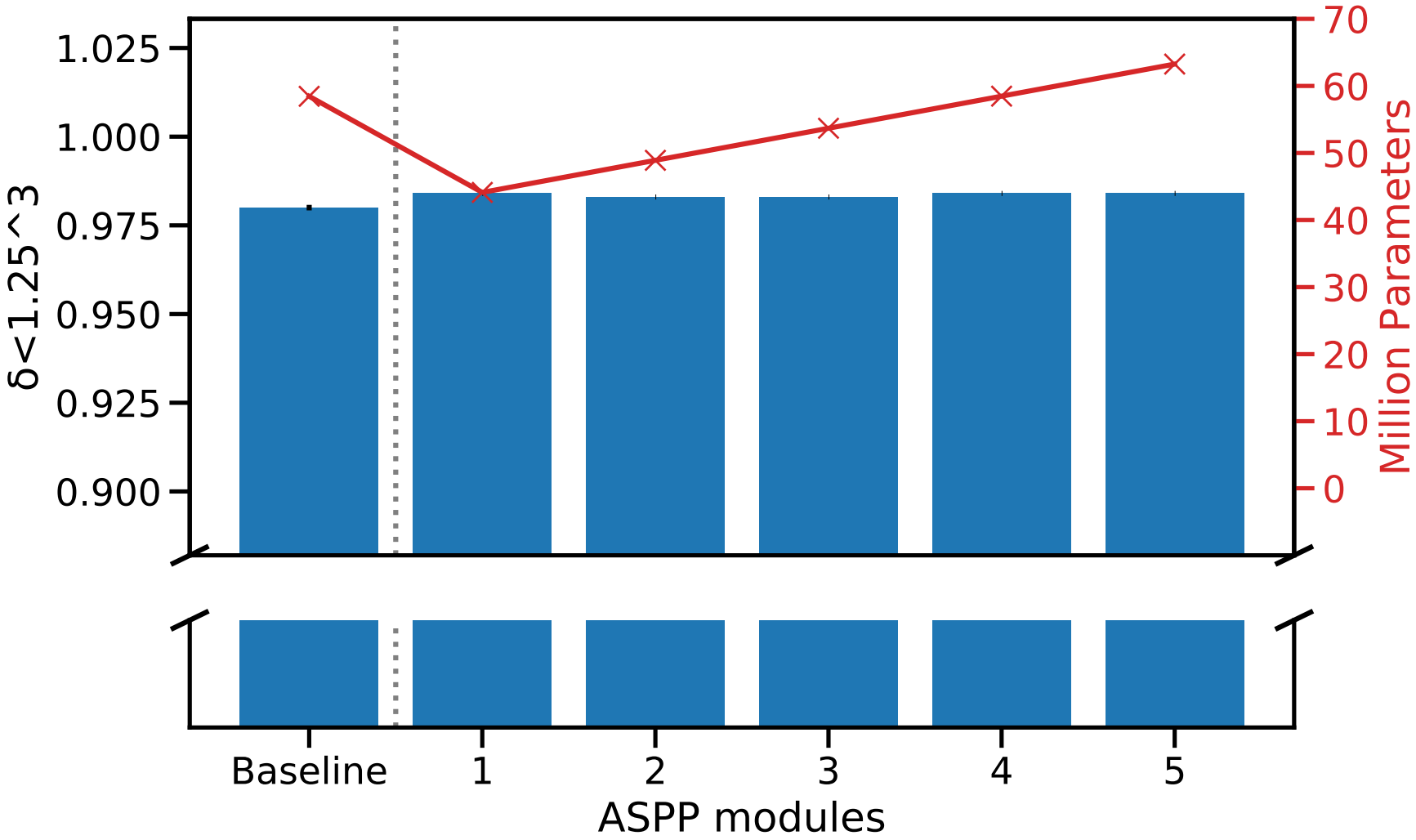
Experiments: Atrous Rates (% inliers 1)



Experiments: Atrous Rates (% inliers 2)

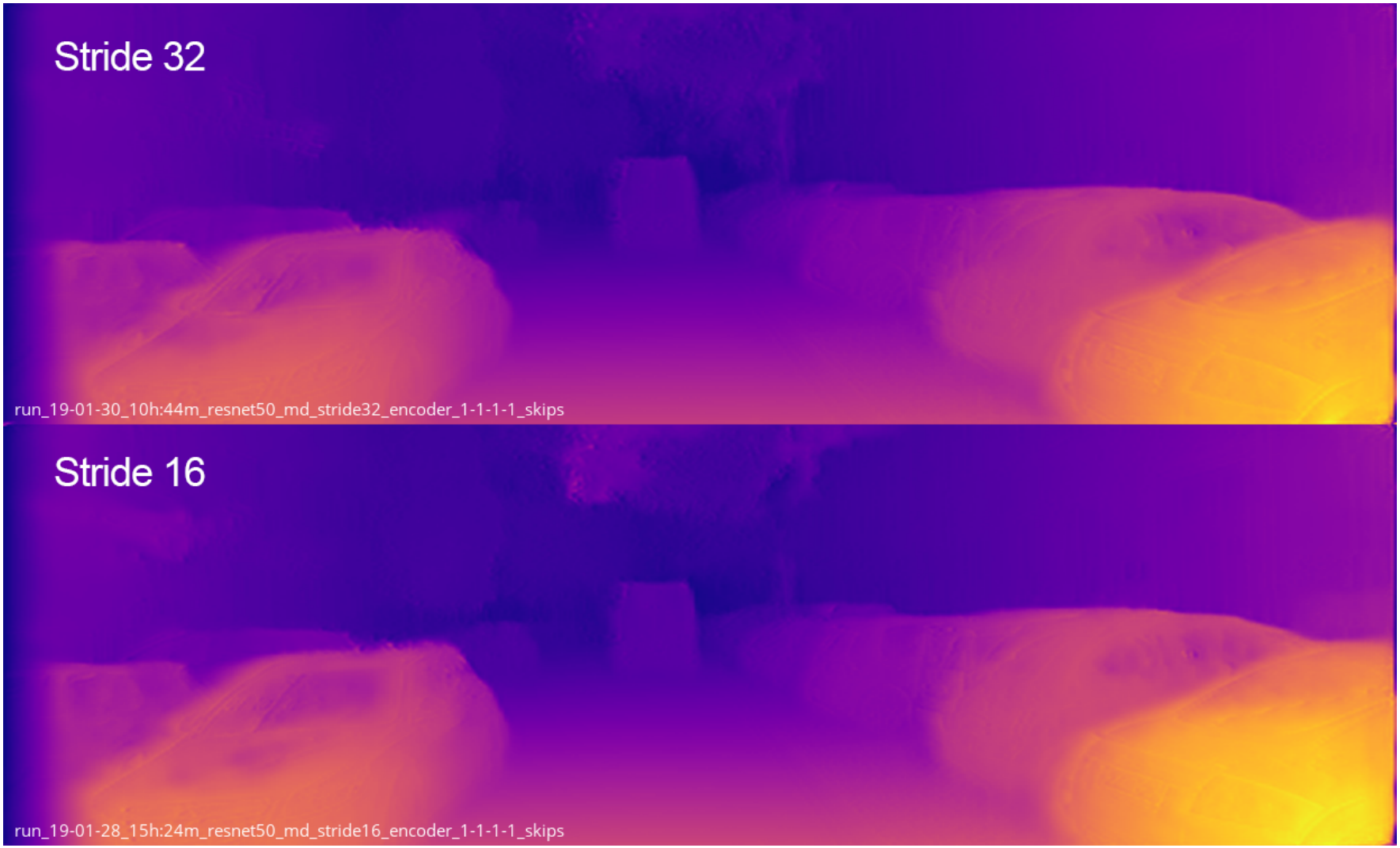


Experiments: Atrous Rates (% inliers 3)



Experiments: Output Stride Comparison

Experiments: Output Stride Comparison



Experiments: ASPP vs. no ASPP (without Skip Connections)